International Journal of Computer Networking,
Wireless and Mobile Communications (IJCNWMC)
ISSN (P): 2250-1568; ISSN (E): 2278-9448
Vol. 8, Issue 1, Jun 2018, 35-44

TRANS STELLAR
•Journal Publications •Research Consultancy

ANALYSIS AND EVALUATION ON ONLINE SHOPS CUSTOMERS THROUGH DATA MINING METHODS

SNEZHANA SULOVA

Department of Computer Science, University of Economics, Varna, Sofia, Bulgaria

ABSTRACT

Customer data analyzers of online stores are important for improving customer service and provide opportunities for greater personalization. In this paper, an approach is proposed for analyzing customer data based on data mining methods. Natural computer language processing technologies are also used to identify customers who are satisfied with the e-shop's service and those who are not. A decision tree is also built that allows for the choice of rational management decisions.

KEYWORDS: Classification, Decision Tree, Machine Learning, Online Shops, Customers & Rapid Miner

Received: Feb 01, 2018; **Accepted:** Feb 22, 2018; **Published:** Jun 14, 2018; **Paper Id.:** IJCNWMCJUN20184

Original Article

INTRODUCTION

In order to improve their business, online traders are forced to carry out different types of analyzes of collecting customer data. One of the major challenges for the development of e-commerce is the application of various Data Mining (DM) methods to research and detect hidden data, the knowledge that was previously unknownbut can be useful to business. Unlike a data retrieval through queries to a database the extraction of knowledge from the data generates unpredictable, previously unknown but potentially useful information. One of the best definitions for DM, which is used in the literature, is made by the researchers from the Massachusetts Institute of Technology (MIT). They define the term as “analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner” [1].

The DM can explore the presence of dependencies, such as associations and sequences, classification, clustering, and forecasting can be done. DM is often defined as a multidisciplinary area because it develops on the basis of the “many techniques from other domains such as statistics, machine learning, pattern recognition, database and data warehouse systems, information retrieval, visualization, algorithms, high-performance computing, and many application domains” [2].

The subjects of learning and application in the present study are machine learning methods. Arthur Samuel defines machine learning as a field of study that gives computers the ability to study without being explicitly programmed [3]. The types of machine learning are Supervised Machine Learning and Unsupervised Machine Learning. In Supervised Machine Learning, a data mining algorithm is used to construct a model, e.g. a classifier. The classifier then undergoes training and the quality of his work is checked. If it is not satisfactory, it undergoes additional training and the process is continued until the desired quality level is reached or until it is found that the

algorithm doesn't work properly with the selected data. Unsupervised Machine Learning is for tasks with descriptive models, e.g. for discovering regularities in purchases. If it has regularity, the model represents it, and it does not require prior knowledge of the analyzed data. Similar tasks are clustered and the search for associative rules [4].

For the purpose of our research, where it is necessary to divide the input data into two or more groups, we use classification, based on groups of the sample, learning data. This classification uses algorithm-based rules such as Naive Bayes (NB) and Support Vector Machines (SVM). For analyzing and presenting the data we use the classification decision tree.

APPROACH FOR ANALYSIS

Our approach to analyzing e-shop customers is based on data collected through a survey with customers and data extracted from the e-shop data base. The main steps of the analysis process are (Figure 1):

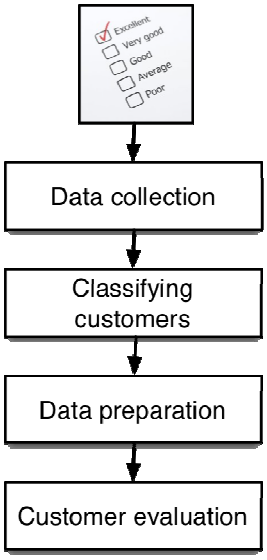

Figure 1: Basic Steps of the Analysis

- Data collection. At this stage, data are collected through a pre-prepared survey. After making a purchase, customers are given the opportunity to write their opinions about the e-shop and whether they are satisfied with the purchase and answer a few short questions.
- Classifying customers into satisfied or not, depending on the polarity of shared opinions about the e-shop. The analysis of opinions is based on existing text-processing technologies, its presentation as a word vector and classification of opinions using supervised machine learning technologies [5]. There are numerous studies in the literature for Opinion mining (OM) and Sentiment analysis (SA). Some of the extensive studies are by Liu [6], K. Ravi and V. Ravi [7]. Taking note of these authors' research and our studies [8], we have found that it is best to use linear SVM or NB [9] [10].
- Data preparation. A data table for further analysis is prepared at this stage. The table for each customer includes data from the evaluation of his/her opinion on the online store and extract data from the database.
- Customer evaluation by building a decision tree, which is a graphical method of choosing an alternative by exploring

consistent and interrelated solutions and their results.

Decision trees are a method of classifying data using a graph in the form of a tree. They are a structure of nodes and links between them. The tree shapes the decision-making process regarding the class to which the classified object is subjected to, by checking logical conditions. The general principle of building a decision tree is based on the calculation of probabilities and the division of subsets. This process recurs recursively until partition stops adding new values, and there can be no breakdown of subsequent subgroups. When building decision trees, it is necessary to choose the features that will separate the sets. The most commonly used algorithms are ID3 [11], C4.5 [12] and Classification and Regression Tree (CART) [13].

The classification algorithm ID3 works by choosing one feature for "best" and dividing the set of objects on it. It calculates the "information gain" indicator based on the concept of entropy of information and selects the feature with the greatest "information gain" that forms a node when building the tree. The node with an entropy of 0 is a leaf and is not subjected to further separation. The node with an entropy greater than 0 is subdivided to the classification of the set of objects. The C4.5 algorithm can be viewed as an extension of ID3. It calculates a "normalized information gain" and thus takes only the significant nodes when constructing the classification tree.

The CART algorithm is designed to build a binary decision tree. To construct each of his nodes, a rule is used that divides the set of objects into two - one for which the rule is executed and another for which it is not executed. This algorithm provides a capability for estimating branch quality, missing value processing, and has a mechanism for optimal tree truncation.

RESULTS AND DISCUSSIONS

Through the survey, we conducted, 181 e-shop customer opinions have been gathered. The data are shown in the table in Figure 2.

Client ID	Opinion	Incame
1001	Overall, my impression of the store is positive. There are many and	1500-2000
1023	One of the best electronic stores I've ordered. I recommend it.	1000-1500
1023	The delivery of the goods is expensive. I would not have ordered ag	<1000
1035	It is very easy to find in the catalog and the goods are described ve	1000-1500
1045	In this shop are quality and good prices. There is a wide variety of it	1000-1500
1202	The delivery of goods is slow. I am not satisfied with the goods rece	<1000

Figure 2: Data from the Survey

Text processing and automated classification of opinions is made by using one of the most popular Data Mining software products – Rapid Minier. It provides an interactive graphical user interface and tools for link analysis, analysis of texts, unstructured data.

To perform the classification, training data on the model is also prepared. The training data are sufficient and well-checked. They are a very important element of the realization of machine self-learning, then the model is validated based on them. The Rapid Miner model of customer opinion classification in polarity is shown in Figure 3.

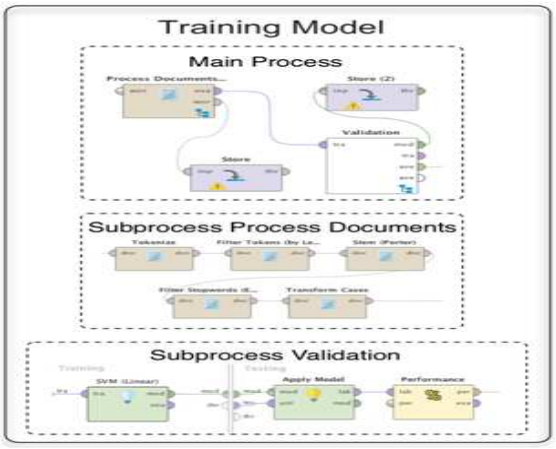

Figure 3: Model for Classifying Customer Opinions using the Svm Method

The Process Documents Operator is used to process the text and allows any opinion to be presented as a vector of words. Text processing includes:

- tokenizes a document;
- Filters tokens based on their length (min chars 4, max chars 25)
- Removes English stop words from a document;
- The Porter stemmer for English words;
- Transforms cases of characters in a document.

When generating the word vector in Rapid Miner, the popular method of evaluating the terms used is the Inverted Document Frequency (TF-IDF), which is a statistic showing how important the word is for a collection of documents or a corpus. TF-IDF increases its value proportionally to the number of word matches, but also considers the frequency of the word in the body because some words tend to appear more often.

Through the created in the figure 3 model, customer reviews of positive and negative opinions are successfully classified. The result is displayed in the form of a table (Figure 4).

Row No.	prediction(label)
1	positive
2	positive
3	negative
4	positive
5	positive
6	negative

Figure 4: Result of An Analysis of Opinions by Rapid miner

Based on the obtained results and with additional data from the e-shop database, a table is constructed from Figure 5. It contains the following fields: Client_ID, Satisfied, Income, Sex, City, Age. Data from the Client_ID, Sex, City, Age fields are retrieved from the database and the other two fields are from the conducted survey. In the Satisfied field, it's entered with a Yes, if, as a result of the automatic analysis of the previous step, the e-shop opinion is considered positive and No if it is

registered as negative.

Client ID	Satisfied	Incame	Sex	City	Age
1001	Yes	1500-2000	Male	Varna	45
1023	Yes	1000-1500	Female	Sofia	40
1023	No	<1000	Male	Veliko Tarnovo	27
1035	Yes	1000-1500	Male	Varna	35
1045	Yes	1000-1500	Male	Varna	43
1202	No	<1000	Female	Sofia	55

Figure 5: A Data Set for Decision Tree Analysis

The data from the table in Figure 5 is analyzed by building a decision tree. The method for machine learning, decision tree, is one of the most popular to solve classification and forecasting tasks [14]. The purpose of building it is to determine the value of a categorical dependent variable. The tree consists of a root node, nodes, and leaves.

As it has already been mentioned, in the decision trees, it is necessary to choose the features on which the sets will be divided. To make an analysis, we build a decision tree with the three algorithms described in the previous paragraph. The trees obtained in the nodes are given in gray. They are attributes which serve as good predictors for our label attribute. Rapid Miner implemented a decision tree which can use different split evaluation criteria. The leaves are the endpoints in blue or red and they show the distribution of the categories of the attribute. In Figure 6 the decision tree that uses the Information gain criterion is shown (Figure 6).

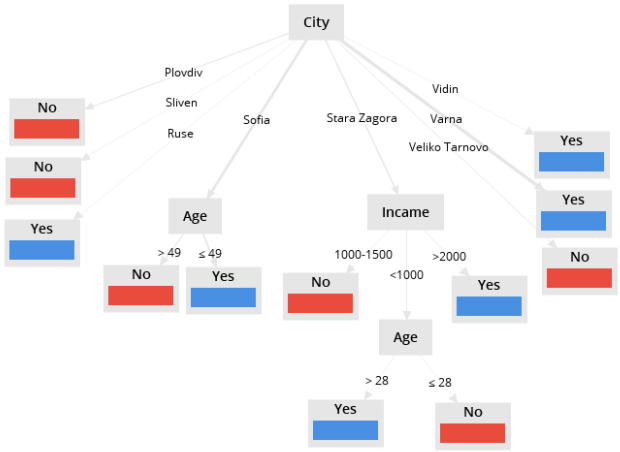

Figure 6: Decision Tree that uses the Information Gain Criterion

Represented in textual form, the decision tree has the following form (Figure 7).

```
City = Plovdiv: No {Yes=0, No=9}
City = Ruse: Yes {Yes=3, No=0}
City = Sliven: No {Yes=0, No=3}
City = Sofia
|   Age > 49: No {Yes=0, No=9}
|   Age ≤ 49: Yes {Yes=15, No=0}
City = Stara Zagora
|   Incame = 1000-1500 : No {Yes=0, No=3}
|   Incame = <1000
|   |   Age > 28: Yes {Yes=3, No=0}
|   |   Age ≤ 28: No {Yes=0, No=3}
|   Incame = >2000: Yes {Yes=3, No=0}
City = Varna: Yes {Yes=30, No=0}
City = Veliko Tarnovo: No {Yes=0, No=6}
City = Vidin: Yes {Yes=3, No=0}
```

Figure 7: Textual form of Decision Tree that uses the Information Gain Criterion

Per the obtained tree, the best indicator of customer satisfaction evaluation is the city. There are no dissatisfied clients from Ruse, Varna, and Vidin, but clients from Plovdiv, Sliven and Veliko Tarnovo are not satisfied. New nodes have been formed in the cities of Sofia and Stara Zagora, respectively by age and income.

By using the Gain ratio criterion, the tree represented in Figure 8 is produced. From the received second decision tree, it is seen that Age is the best indicator of whether a customer is satisfied or not with the store's service. If the age is greater than 50, we see that customers are not satisfied with the e-shop. If, however, the age is less than or equal to 50, then gender becomes an indicator of satisfaction. If the gender is male, then it turns out that the city is important and a new knot is formed. For one of the cities - Stara Zagora there is a next node age.

Figure 8: Decision Tree that uses the Gain Ratio Criterion

Presented in textual form, the decision tree has the following form (Figure9).

```
Age > 50: No {Yes=0, No=24}
Age ≤ 50
|   Sex = Female: Yes {Yes=68, No=0}
|   Sex = Male
|   |   City = Plovdiv: No {Yes=0, No=18}
|   |   City = Sliven: No {Yes=0, No=6}
|   |   City = Sofia: Yes {Yes=13, No=0}
|   |   City = Stara Zagora
|   |   |   Age > 28
|   |   |   |   Age > 41: No {Yes=0, No=6}
|   |   |   |   Age ≤ 41: Yes {Yes=7, No=0}
|   |   |   Age ≤ 28: No {Yes=0, No=6}
|   |   City = Varna: Yes {Yes=26, No=0}
|   |   City = Veliko Tarnovo: No {Yes=0, No=6}
```

Figure 9: Textual form of Decision Tree that uses the Gain Ratio Criterion

A tree was also built using the Gini index criterion on figure 10.

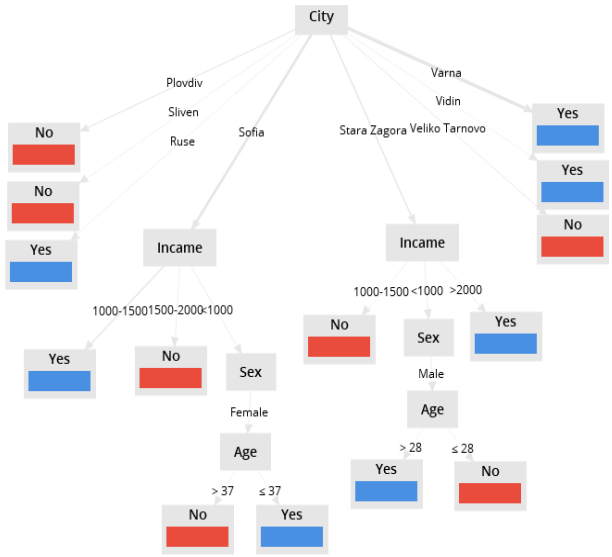

Figure 10: Decision Tree that uses the Gini Index Criterion

Presented in textual form the decision tree has the following form (figure 11).

```
City = Plovdiv: No {Yes=0, No=9}
City = Ruse: Yes {Yes=3, No=0}
City = Sliven: No {Yes=0, No=3}
City = Sofia
|   Incame = 1000-1500 : Yes {Yes=12, No=0}
|   Incame = 1500-2000 : No {Yes=0, No=6}
|   Incame = <1000
|   |   Sex = Female
|   |   |   Age > 37: No {Yes=0, No=3}
|   |   |   Age ≤ 37: Yes {Yes=3, No=0}
City = Stara Zagora
|   Incame = 1000-1500 : No {Yes=0, No=3}
|   Incame = <1000
|   |   Sex = Male
|   |   |   Age > 28: Yes {Yes=3, No=0}
|   |   |   Age ≤ 28: No {Yes=0, No=3}
|   Incame = >2000: Yes {Yes=3, No=0}
City = Varna: Yes {Yes=30, No=0}
City = Veliko Tarnovo: No {Yes=0, No=6}
City = Vidin: Yes {Yes=3, No=0}
```

Figure 11: Textual form of Decision Tree that uses the Gini Index Criterion

From the tree obtained using the Gini index again the city was chosen as the best indicator for satisfaction assessment. The results are similar to the first decision tree. The unsatisfied clients are from Plovdiv, Sliven and Veliko Tarnovo, and in the cities of Sofia and Stara Zagora anode is formed by income. For customers with a certain income, additional units have been obtained by gender and age.

For all three decision trees built, the Performance operator has been used. Accuracy is calculated for the first two classification trees at 88.89, and for the latter it is 81.48.

This analysis gives us reason to make the following conclusions for our e-shop customers:

- The residence is of great importance for satisfaction, which means that, as the souvenirs are essential goods for each region, frustration can be due to problems with the delivery of goods to these cities;
- Older customers are not satisfied, probably the goods offered, are items that are either not suitable for them or are expensive and they cannot buy them;
- Gender is also an indicator that matters, the analysis show that, overall, women are more satisfied;
- Income can be used as an indicator, but in combination with the city, because there is a different city satisfaction with the same income.

CONCLUSION

In this paper, an approach is proposed for analyzing customer data based on data mining methods. The approach includes processing of survey data and data from the e-shop data base. Text mining technologies and classification through decision trees are used. The suggested approach has been tested with the Rapid Miner software and the results show that it can be successfully used by online marketers to make adequate solutions related to the customer relationship management process.

REFERENCES

1. *D. Hand, H. Mannila, and P. Smyth, "Principles of Data Mining", MIT Press, 2001.*
2. *J. Han, M. Kamber, J. Pei, "Data Mining Concepts and Techniques", Third Edition, Morgan Kaufmann Publishers, 2012.*

A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers", IBM Journal 3, 211-229, 1959.

3. *A. Barsegyan, M. S. Kupriyanov, I. I. Kholod, "Analysis of data and processes", 3rd ed., (Russian), BKhV - Sankt-Peterburg, 2009.*
4. *V. Gupta, G. S. Lehal, "A Survey of Text Mining Techniques and Applications", Journal of emerging technologies in web intelligence, vol. 1, no. 1, pp. 60-76, 2009.*
5. *Liu, "Sentiment Analysis and Opinion Mining", Morgan & Claypool Publishers, 2012.*
6. *K. Ravi, V. Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications", Knowledge-Based Systems, vol. 89, pp. 14-46, 2015.*
7. *S. Sulova, "An approach for automatic analysis of online Store product and services reviews", Izvestia, Journal of Varna University of Economics, vol. 60, no 4, pp 455-467, 2016.*
8. *P. Singh, M. Husain, "Methodological study of opinion mining and sentiment analysis techniques", International Journal on Soft Computing (IJSC), vol. 5, no 1, p. 11-21, 2014.*

9. *R. Verma, Kiranjyoti, "Opinion Mining and Analysis of the Techniques for User Generated Content (UGC)" – International Journal of Advanced Research in Computer Science and Software Engineering, no 5, pp. 438-441, 2005.*
10. *J.R. Quinlan, "Induction of decision trees", Machine Learning 1, pp. 81-106, 1986.*
11. *Sao, Nikita, and Ravi Mishra. "Video Shot Boundary Detection based on Nodal Analysis of Graph Theoritic Approach." (2014).*
12. *S. L. Salzberg, "C4.5: Programs for Machine Learning by J. Ross Quinlan", Morgan Kaufmann Publishers, Inc., vol. 16, issue 3, pp 235–240, 1994.*
13. *L. Breiman, J. Friedman. C. J. Stone and R. A. Olshen, "Classification and regression trees", CRC Press, 1984.*
14. *M. Kantardzic, M, "Data mining: concepts, models, methods, and algorithms", John Wiley & Sons, 2011.*